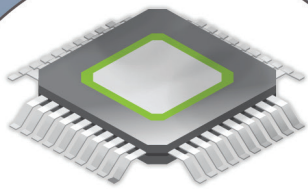
CHAPTER 01

Computer Architecture

This laptop has a processor that runs up to 4.5 GHz, 8 GB of memory and 512 GB of storage.

Well, I understand that storage is used for saving files and apps, but what about the processor and memory?



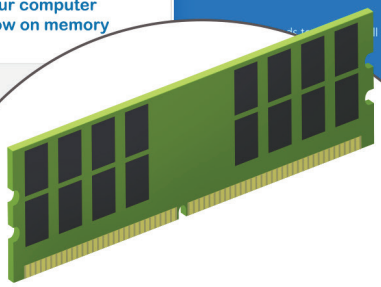
The processor is the "brain" of the computer. It executes instructions and performs calculations.

Yes. And the memory is where data and instructions are temporarily stored so that the processor can access them quickly.

I see, so these calculations are done by the processor!

So, is the memory important for multitasking?

Your computer is low on memory



Yes. Without sufficient memory, applications will frequently need to load and unload data from storage, hence slowing down the computer.

There's so much to learn about the parts of a computer and their functions.

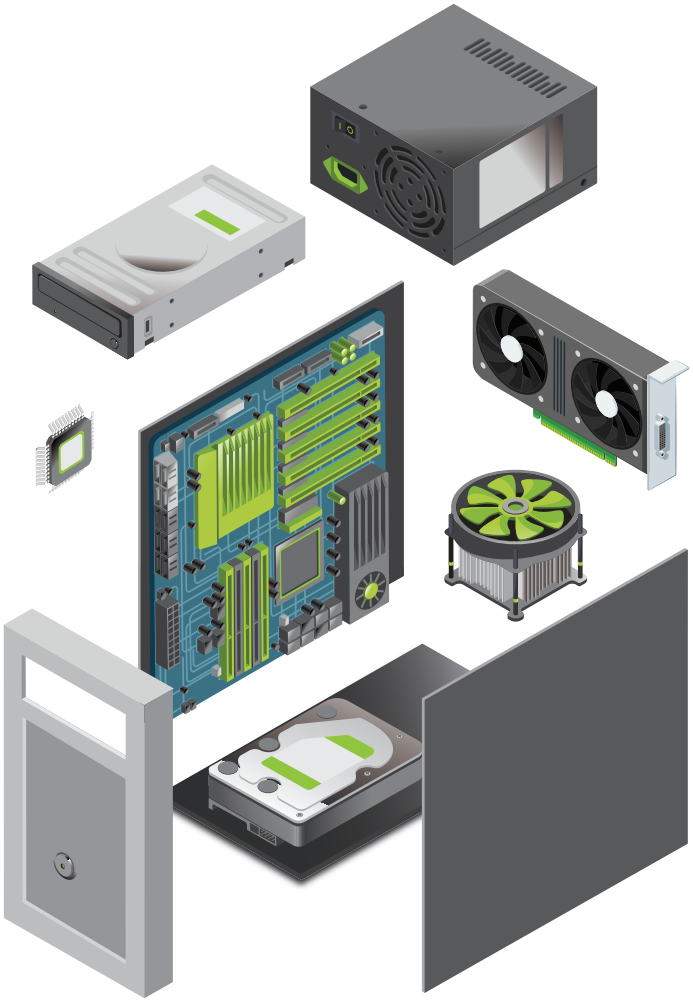
Thank you for answering our questions!

No problem. Is there anything else do you want to know?

Yes, could you tell us more about...

1.1 Introduction to Computer Architecture

A **computer** (or computer system) is a device that receives and processes data according to a set of instructions, and produces the processed data as a result. **Computer architecture** describes how a computer is designed and built to function. It also includes how the various parts of a computer are designed, organised and connected.



KEY TERMS

Computer
A device that receives and processes data according to a set of instructions and produces the processed data.

Computer architecture
A description of how a computer is designed and built to function, including how its various parts are designed, organised and connected.

Data (singular: datum)
Information that is used in a computer program.

DID YOU KNOW?

There are many ways to build a computer. While most modern computers run on electricity, early computers did not use electricity at all. Instead, they relied on mechanisms such as cranks, gears, pulleys and levers. These mechanical computers used a completely different computer architecture from the computers we are familiar with today.

For example, Figure 1.1 shows a toy called the Dig-Comp II that uses a set of switches to represent data. These switches are connected to ramps. As marbles roll down the ramp, they trigger the switches in different ways. Because of this, the toy is able to perform computations such as addition, functioning as a mechanical computer that works without electricity.

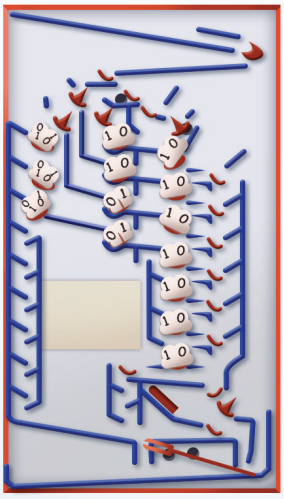


Figure 1.1 Digi-Comp II

1.2 Units of Data

LEARNING OUTCOMES

1.1.1 *Perform calculations using bits, bytes, kilobytes, kibibytes, megabytes, mebibytes, gigabytes, gibibytes, terabytes, tebibytes, petabytes and pebibytes.*

Since a computer's function is to receive, process and produce data, an important aspect of computer architecture is to define how data is being represented.

Modern computers are digital devices evolved from machines that perform calculations on numerical data represented in binary form. The smallest unit of data is a **bit**. A bit is a binary digit that takes on a value of either 0 or 1. A computer stores and processes all data using binary numbers that consist of these digits.

A single bit would be too simple to represent complex data, so we usually talk about data being represented as bytes instead. A **byte** is a unit of data made up of eight bits.

Table 1.1 summarises some units of measurement of data in order of increasing size.

KEY TERMS

Bit
A binary digit that takes on a value of either 0 or 1.

Byte
A binary number made up of eight bits.

Name of unit	Symbol	Size in bytes
kilobyte	kB	1,000
kibibyte	KiB	1,024
megabyte	MB	$1{,}000^2 = 1{,}000{,}000$
mebibyte	MiB	$1{,}024^2 = 1{,}048{,}576$
gigabyte	GB	$1{,}000^3 = 1{,}000{,}000{,}000$
gibibyte	GiB	$1{,}024^3 = 1{,}073{,}741{,}824$
terabyte	TB	$1{,}000^4 = 1{,}000{,}000{,}000{,}000$
tebibyte	TiB	$1{,}024^4 = 1{,}099{,}511{,}627{,}776$
petabyte	PB	$1{,}000^5 = 1{,}000{,}000{,}000{,}000{,}000$
pebibyte	PiB	$1{,}024^5 = 1{,}125{,}899{,}906{,}842{,}624$

Table 1.1 Units of measurement of data

DID YOU KNOW?

In the past, units of data such as "kilobyte", "megabyte" and "gigabyte" were based on powers of 1,024 (or 2^{10}) instead of the standard powers of 1,000 used in the International System of Units (SI). However, since 1998, multiple standards organisations have agreed that SI prefixes (i.e., "kilo", "mega", "giga", etc.) should follow their standard meanings while new binary prefixes (i.e., "kibi", "mebi", "gibi", etc.) would be used to represent powers of 1,024.

Despite this change, you may still encounter the use of "kilobyte" and "megabyte" to represent 1,024 bytes and $1{,}024^2$ bytes respectively, instead of the correct units "kibibyte" and "mebibyte".

QUICK CHECK 1.2

1. 1 kilobyte is the same size as 1,000 bytes. True or false?

2. Arrange the following units in increasing size: bit, byte, gibibyte, gigabyte, kibibyte, kilobyte

3. Convert the following amounts of data into bytes:
a) 2,026 kB **b)** 19 GiB **c)** 65 MB

4. Convert the following amounts of data into bits:
a) 2,026 kB **b)** 19 GiB **c)** 65 MB

1.3 Components of a Computer System

LEARNING OUTCOMES

1.1.2 *Describe the function of key components of a computer system: its processor, main memory and secondary storage.*

1.1.3 *Describe the function of data and address buses in reading from and writing to memory.*

1.1.4 *Describe different input/output interfaces (USB, HDMI and PCI Express) in terms of typical applications, connectors and speed.*

1.1.5 *Describe the use of magnetic, optical and solid-state media for secondary storage in terms of durability, portability, typical capacities, cost and speed.*

Most computer parts can be organised into the following roles in Table 1.2. Figure 1.2 shows where the parts performing these roles may be found inside the case of a desktop computer.

DID YOU KNOW?

Instead of having separate parts, the key components of a computer in mobile devices such as smartphones and laptops are combined into a single part called a System on a Chip (SoC). This has the benefits of smaller size, lower power consumption and, often, improved performance. However, SoCs are also more complex and require specialised expertise to design and manufacture.

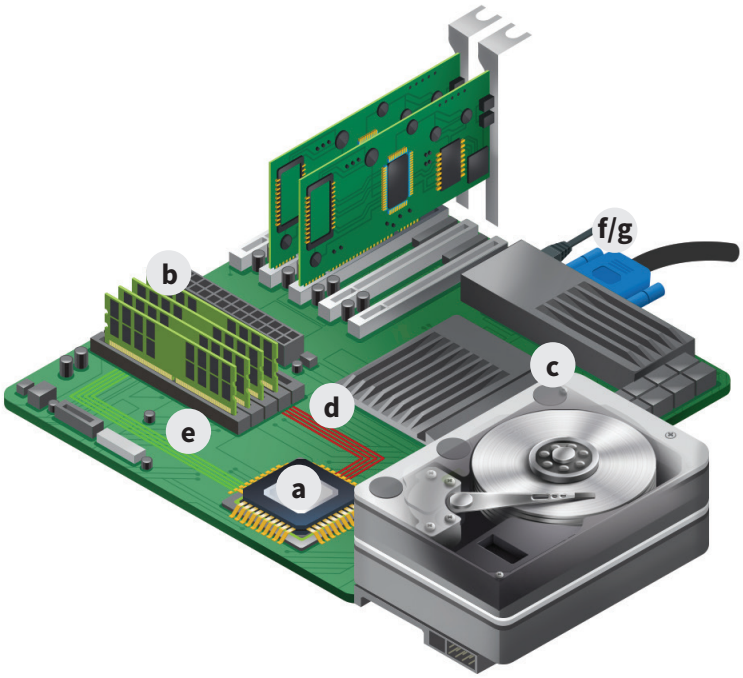


Figure 1.2 Components in a desktop computer case

Role	Description
(a) Processor	Processes data and follows instructions
(b) Memory	Stores data and instructions temporarily for immediate use by the processor
(c) Secondary storage	Stores large amounts of data that will not be lost when the power supply is interrupted
(d) Data bus	Transports data between memory and processor; bi-directional
(e) Address bus	Transports required memory location from processor to memory; uni-directional
(f) Input	Data or instructions that the computer receives
(g) Output	Intermediate or final results produced by the computer; usually in the form of processed data

Table 1.2 Inside the computer

1.3.1 Processor

You have learnt that computers need to process data and follow instructions. The computer part that does this is called the **processor** or **central processing unit (CPU)**. It is usually a complex circuit made of many components compressed into a square or rectangular package. You will learn more about the components that make up this circuit in Chapter 3.

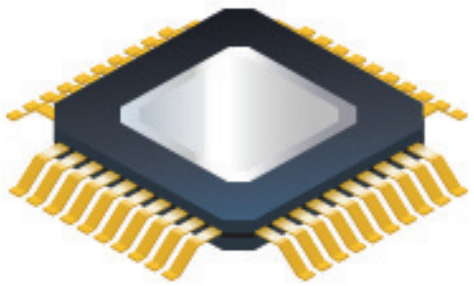


Figure 1.3 A processor package

KEY TERMS

Central processing unit (or processor)
The part of the computer that processes data and follows instructions.

Execute
To follow or perform an instruction.

Software
A set of instructions to perform specific tasks on a computer.

When a processor 'runs' or '**executes**' instructions, it follows or performs instructions. A series of instructions is called a program, or simply **software**.

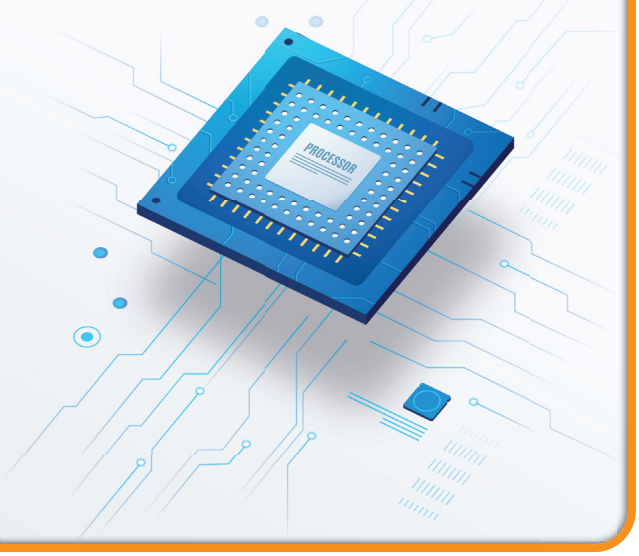
Often, a processor's speed is described by the number of instructions that the processor can perform in one second. For instance, a 1 MHz (megahertz) processor can perform one million instructions per second while a 1 GHz (gigahertz) processor can perform one billion instructions per second. In general, the larger the number and unit, the faster and more powerful the processor.

There are also "multi-core" processors that contain multiple processing units inside a single package. For instance, a "dual-core" package has two such processing units while a "quad-core" package has four. These "multi-core" processors can perform more than one instruction at the same time, and thus are more powerful than "single-core" processors.

DID YOU KNOW?

Many computers also include a specialised computer part called a Graphics Processing Unit (GPU) that is more efficient at performing the calculations needed to produce images than a CPU. By relieving the CPU of tasks related to 2D and 3D graphics, GPUs help to improve the overall performance of a system.

Since graphics often involve performing many calculations simultaneously, GPUs have also become useful in fields such as Artificial Intelligence that require performing repetitive calculations on a large amount of data. You will learn more about this topic in Chapter 14.



1.3.2 Main Memory

Whenever a computer needs to store data, it uses **memory.**

There are different types of memory. Random access memory (RAM) or main memory is a type of memory where data and instructions are stored temporarily so that they can be quickly accessed by the processor when needed. For instance, when an application is started, its instructions may be loaded into the RAM. Data stored on the RAM can be easily changed and is also **volatile**, which means that it is lost once the power supply to the computer is interrupted.

The word "memory" usually refers to RAM instead of other types of memory. Most desktop computers have RAM located on removable cards so that the amount of RAM in the computer can be upgraded easily. Other kinds of computers, however, may have only a fixed amount of RAM that cannot be changed.

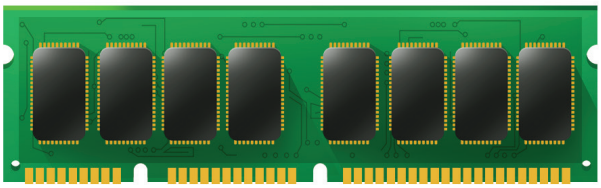


Figure 1.4 Example of a RAM card

DID YOU KNOW?

Read-only memory (ROM) is another type of memory that stores data and instructions that rarely need to change or would be needed for a computer to start up. Data stored on ROM cannot be easily changed and it is retained even when the power supply is switched off. This makes it ideal for storing critical instructions that are required to start the computer before other data can be loaded into the RAM.

Physically, memory exists in different forms, but in general, it consists of many switches arranged in a fixed order. Each switch can store one bit of data based on whether it is ON or OFF (i.e., ON = 1, OFF = 0). Since a byte consists of eight bits, we usually look at eight consecutive switches at a time, as shown in Figure 1.5.

In Figure 1.5, big physical switches are used to represent memory, but in most computers, memory is made up of very small electronic switches that take up much less space.

Usually, the position of each byte is represented by a number called an **address**. This is a number that allows a computer to quickly find those switches again if it needs to read or change the data that is stored. This is similar to how a unit number is used to identify the location of each residence in an HDB block.

KEY TERMS

Address
A number that is used to locate a byte in memory.

Memory (Main Memory)
A device that is used to store data and instructions temporarily for immediate use by the processor.

Volatile
Lost when the power supply is interrupted.

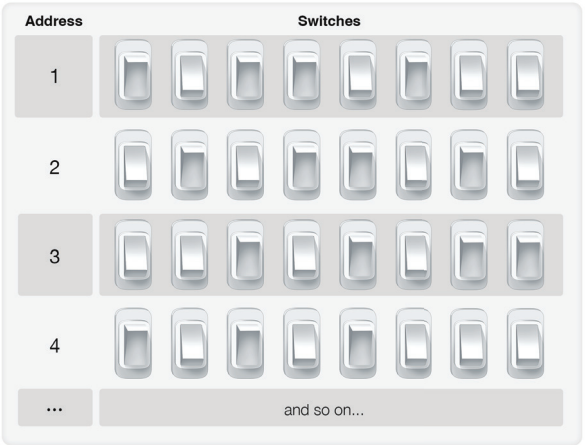


Figure 1.5 Memory as a collection of switches in a fixed order

1.3.3 Secondary Storage

Secondary storage is a way of storing large amounts of data that will not be lost when the power supply is interrupted. Compared to RAM, secondary storage is usually cheaper and able to store much more data. It is also non-volatile, so the data that is stored remains there even without a power supply. This makes secondary storage ideal for physically transporting data from one computer to another. On the other hand, secondary storage is usually much slower in speed compared to RAM.

The word "storage" usually refers to secondary storage.

The processor usually does not access data in secondary storage directly. Instead, any data in secondary storage that the processor needs might be copied to the RAM first.

KEY TERMS

Secondary storage
A way of storing large amounts of data that will not be lost when the power supply is interrupted.

Hard disk (or hard drive)
Secondary storage where data is stored on rigid rotating disks coated with a magnetic material.

1.3.3.1 Choosing Storage Media for Secondary Storage

There are many types of storage media available. When deciding on the best type of storage media to use, these factors should be considered:

Figure 1.6 Factors to consider when choosing storage media

1.3.3.2 Types of Storage Media

Table 1.3 summarises three main types of storage media in terms of durability, portability, typical capacities, cost and speed.

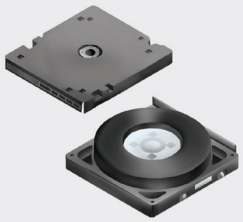
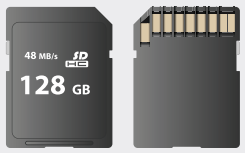
Type	Magnetic	Optical	Solid-State
Description	Data is stored on a magnetic material that can be read or written by a magnetic "head" Example: hard disk  Example: magnetic tape 	Data is stored as very small pits or indentations that can be read or written by a laser Example: digital versatile disc (DVD) 	Data is stored in electronic circuits that have no moving parts Example: memory cards  Example: solid-state drive 
Durability	More vulnerable to damage from • magnetic fields, • heat, • impact, and • natural deterioration over time	Vulnerable to damage from • scratches, and • natural deterioration over time More resistant to • heat • impact	Most durable; Most resistant to • impact, and • temperature changes
Portability	Heavier and bulkier than optical and solid-state media	Portable due to small size and light weight	Portable due to small size and light weight
Typical Capacities	Up to TBs of data	Up to GBs of data	Up to TBs of data
Cost per GB	Lowest	Lower cost than solid-state media but higher cost than magnetic media	Highest
Speed	Slower than solid-state media	Slower than solid-state media	Faster than magnetic and optical media

Table 1.3 Three main types of storage media

1.3.4 Data and Address Buses

Besides storing data, computers also use buses to transport data from one part of the computer to another. A bus is a collection of wires that serves as a "highway" for data to travel on. It can be made of either physical wires or conductive lines printed on a circuit board.

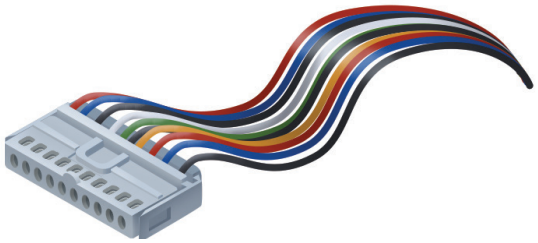


Figure 1.7 Example of a computer bus made of physical wires

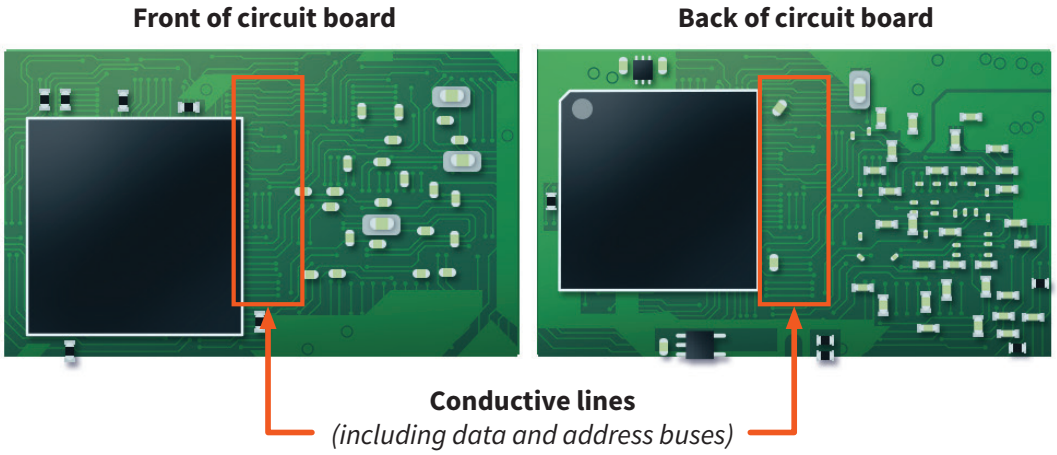


Figure 1.8 Example of computer buses made of conductive lines

KEY TERMS

Bi-directional
Able to work in two directions, to and fro.

Data bus
A bus that is used to transport data between memory and the processor.

Address bus
A bus that is used to specify memory address information.

Uni-directional
Able to work in one direction only.

Two important buses that transport data between the processor and memory parts of a computer are the data bus and the address bus:

1. The **data bus** transports data that is going to be processed to the CPU, and transports data that has already been processed from the CPU. The data bus is **bi-directional** because data can be sent in both directions between the processor and memory.

2. The **address bus** specifies memory address information. When the processor reads from or writes to the memory (RAM), the relevant address information is provided on the address bus. The address bus is **uni-directional** because address information is always sent in one direction only, that is, from the processor to the memory.

For instance, to read data from the memory, the processor requests the data, and the address bus transports the requested data's address to the RAM. A copy of the requested data is then sent from the RAM back to the processor via the data bus. This is illustrated in Figure 1.9, which uses bits for data and simple numbers for addresses.

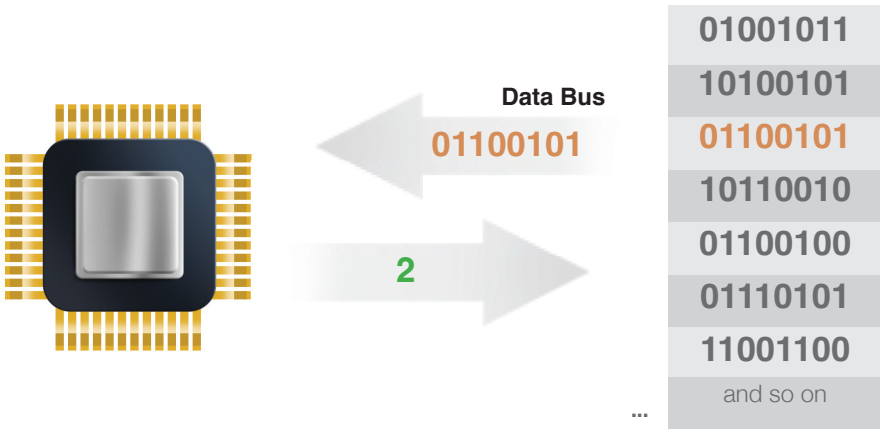


Figure 1.9 How the data bus and address bus are used to read from memory

When writing data to the memory, the processor uses both the data bus and address bus at the same time to transport the data for writing as well as the destination address to the RAM. The RAM then sets the switches at the destination address according to the data received via the data bus, as illustrated in Figure 1.10.

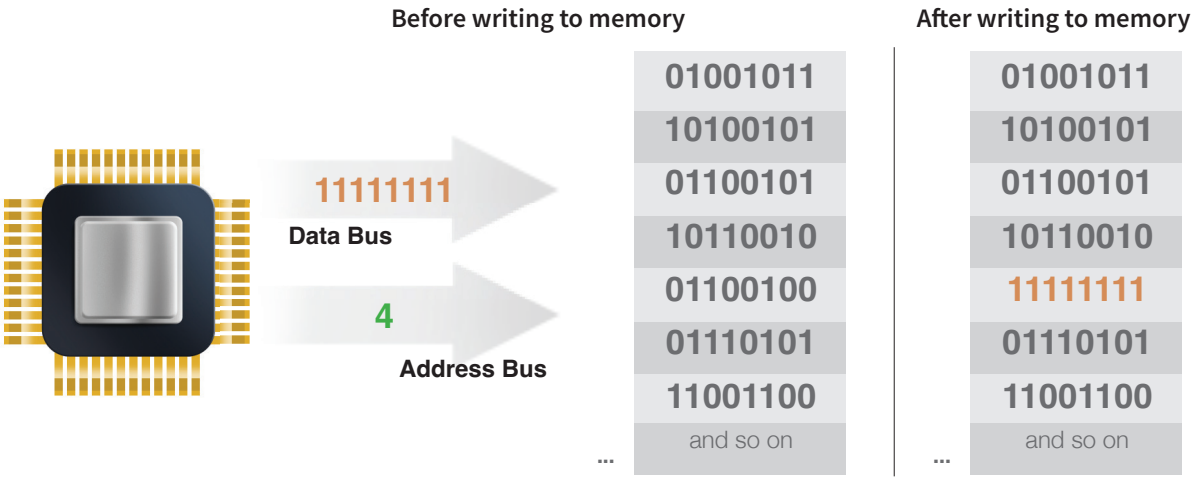


Figure 1.10 How the data bus and address bus are used to write to memory

As shown by the arrows in Figure 1.9 and Figure 1.10, while the address bus is always used to transport information from the processor to the memory, the data bus can transport information in either direction.

1.3.5 Input and Output Interfaces

In computer architecture, **input** refers to data or instructions that the computer receives for processing while **output** refers to any intermediate or final results produced by the computer in the form of processed data. Examples of input are words entered using a keyboard, pictures taken by a digital camera, and movement instructions entered using a mouse. Examples of output are images displayed on a screen, sounds played on a speaker and even sculptures printed using a 3D printer.

Often, input is received from, and output is sent to **hardware** devices:

- An **input device** is a hardware device that allows users to enter data and instructions into a computer. Examples of input devices are keyboards, mice, scanners, touch screens and microphones.
- An **output device** is a hardware device used to display, project or print processed data from a computer so it can be used or understood by people using the computer. Examples of output devices are monitors, speakers and printers.

A computer usually has multiple input and output devices connected to it, as seen in Figure 1.11.

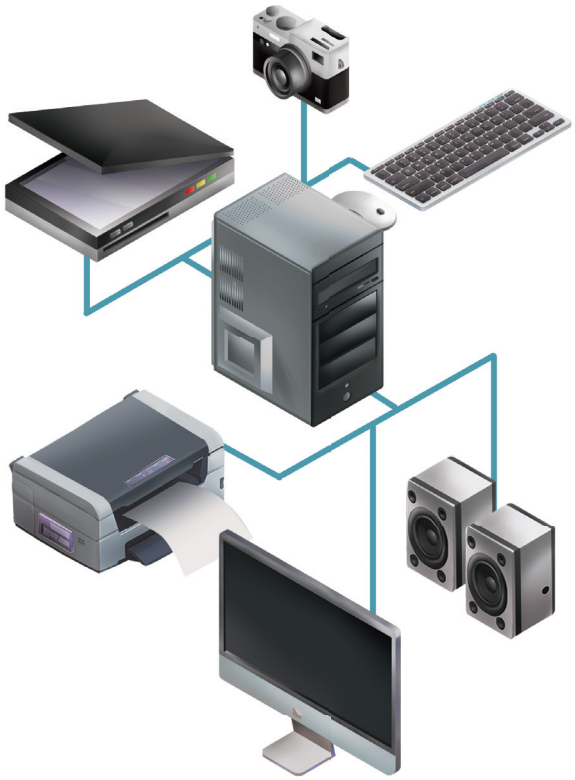


Figure 1.11 Examples of connected input and output devices

KEY TERMS

Input (computer architecture)
Data or instructions that the computer receives for processing

Output (computer architecture)
Intermediate or final results produced by the computer in the form of processed data

Hardware
Physical components of a computer

Input device
A hardware device that allows users to enter data and instructions into a computer

Output device
A hardware device used to display, project or print processed data from a computer

The method by which a computer connects to an input or output device is called an input/output interface. Different input/output interfaces have varied uses, physical connectors and maximum data transmission speeds as new versions of each interface are introduced over time. Table 1.4 summarises this information for three common input/output interfaces:

Interface	Typical Applications	Connectors	Maximum Speed
Universal Serial Bus (USB)	For powering and/or communicating with external devices (e.g., printers, mice, keyboards)	USB Type A  USB Micro  USB Type B  USB Type C  USB Mini 	USB 2.0: 480 Mbit/s USB 3.2: 20 Gbit/s USB4: 80 Gbit/s

Interface	Typical Applications	Connectors	Maximum Speed
High-Definition Multimedia Interface (HDMI)	For delivering audio/video data to compatible devices (e.g., monitors, TVs)	HDMI Standard  HDMI Mini  HDMI Micro 	HDMI 1.3-1.4b: 10.2 Gbit/s HDMI 2.0-2.0b: 18 Gbit/s HDMI 2.1: 48 Gbit/s
Peripheral Component Interconnect Express (PCI Express)	For communicating with internal expansion cards (e.g., graphics cards)	PCI-e x1  PCI-e x4  PCI-e x8  PCI-e x16 	Increases with number of lanes (up to x16) Max speed per lane PCI Express 5.0: 4 GB/s PCI Express 6.0: 8 GB/s PCI Express 7.0: 16 GB/s

Table 1.4 Three common input/output interfaces

Unlike USB and HDMI which typically have connectors on the side or back of a computer, PCI Express connectors are typically located inside a computer on its **motherboard**. Each connector has a number of **lanes** for transferring data and the number of lanes determines the length of the connector as well as the maximum speed for that connector. By convention, the number of lanes is described using the abbreviation "x" followed by the number of lanes (e.g., x1, x4, x8 and x16).

While it is not necessary to memorise the maximum speed of each interface, it is useful to recognise that the maximum speed of each interface roughly doubles or more with each major revision and that the internal PCI Express interface is generally faster than the external USB and HDMI interfaces. By convention, the speed of PCI Express lanes is expressed in GB/s instead of gigabits per second (Gbit/s) and that 8 Gbit/s = 1 GB/s.

KEY TERMS

Lane (PCI Express)
An interface for transferring data between a computer and an expansion card.

Motherboard
The main circuit board in a computer that connects all the components together.

QUICK CHECK 1.3

1. Decide if each of the following statements on memory and storage are true or false.
- **a)** Both memory and storage may contain data.
- **b)** The contents of memory are retained when a computer restarts.
- **c)** The contents of storage are retained when a computer restarts.
- **d)** The typical cost per GB is cheaper for memory than for storage.
- **e)** The typical capacity is higher for memory than for storage.

2. For each of the following situations, suggest an appropriate secondary storage medium that should be used and give a reason for your choice:
- **a)** Storing terabytes of video files for backup
- **b)** Keeping an emergency copy of some important files in your wallet

3. The following table shows the first 4 bytes of a computer's memory:

Address	Contents
0	00000000
1	11110000
2	00110011
3	00001111

The processor then executes an instruction to write the 8 bits 11111111 to address 3.
- **a)** To execute the instruction, describe what information must be sent on the address bus and in what direction.
- **b)** Show the resulting first 4 bytes of memory after the instruction is executed.

4. Memory addresses can also be read from or written to memory as data. True or false?

5. Which of the three I/O interfaces (USB, HDMI and PCI Express) are suitable for **directly** connecting:
- **a)** a wireless network adapter?
- **b)** an internal sound card?
- **c)** a pair of audio speakers?
- **d)** a microphone?

REVIEW QUESTION

1. Describe the function of each of the following components:
- **a)** Processor
- **b)** Main memory
- **c)** Secondary storage

2. Calculate the number of bits (not bytes) in the following amounts of data:
- **a)** 8 TB
- **b)** 0.125 MB
- **c)** 0.125 MiB

3. The position of each byte in memory is called its address.
- **a)** Calculate the number of addresses needed for 64 KiB of memory.
- **b)** The memory in a computer has 256 possible addresses. Calculate the total size of the computer's memory in bits (not bytes).

4. A hard disk is used as secondary storage for a laptop.
- **a)** A data bus connects the laptop's processor directly to its hard disk. True or false?
- **b)** The laptop's owner replaces the hard disk with a solid-state drive. Identify two improvements that the owner is likely to experience.

ANSWER

Pg. 6-Quick Check 1.2

1. *True. However, "kilobyte" is still often confused with "kibibyte", which is 1,024 bytes.*

2. *bit, byte, kilobyte, kibibyte, gigabyte, gibibyte*

3. *Show workings for each conversion.*
   a) *2,026 kB = 2,026 × (1,000 bytes) = 2,026,000 bytes*
   b) *19 GiB = 19 × ($1,024^3$ bytes) = 20,401,094,656 bytes*
   c) *65 MB = 65 × ($1,000^2$ bytes) = 65,000,000 bytes*

4. *Note that the answers below are just the answers for Q3 multiplied by 8 because 1 byte is the same as 8 bits.*
   a) *2,026 kB = 2,026 × (1,000 × 8 bits) = 16,208,000 bits*
   b) *19 GiB = 19 × ($1,024^3$ × 8 bits) = 163,208,757,248 bits*
   c) *65 MB = 65 × ($1,000^2$ × 8 bits) = 520,000,000 bits*

Pg. 16-Quick Check 1.3

1. *State true/false for each part (explanations are not required).*
   a) *True.* **b)** *False.* **c)** *True.*
   d) *False. (Storage usually is cheaper.)*
   e) *False. (Storage usually has higher capacity.)*

2. *For each part, give a suggestion and state a reason.*
   a) *A hard disk. It is the most common and affordable device with storage capacity available in terabytes.*
   b) *A memory card. It is the only device small and flat enough to fit in a wallet. A memory card would also be more durable and resistant to impacts that could occur in a wallet.*

3. *Answers as follows:*
   a) *The address 3 must be sent on the address bus. It is sent from the processor to memory.*
   b) *Only the contents at address 3 is changed:*

Address	Contents
0	00000000
1	11110000
2	00110011
3	11111111

4. *True. Memory addresses are just numbers so they can also be treated as data.*

5. *Suitable I/O interfaces:*
   a) *USB and PCI Express* **b)** *PCI Express only*
   c) *USB and HDMI* **d)** *USB only*

ANSWER

Pg. 16-Review Question

1. *Based on definitions of each key term:*
   - **a)** *Processor: processes data and follows instructions.*
   - **b)** *Main memory: stores data and instructions temporarily for immediate use by the processor.*
   - **c)** *Secondary storage: a way of storing large amounts of data that will not be lost when power supply is interrupted.*

2. *Show workings for each conversion.*
   - **a)** *8 × 1,000 × 1,000 × 1,000 × 1,000 × 8 = 64,000,000,000,000 bits*
   - **b)** *0.125 × 1,000 × 1,000 × 8 = 1,000,000 bits*
   - **c)** *0.125 × 1,024 × 1,024 × 8 = 1,048,576 bits*

3. *Show relevant workings for each part.*
   - **a)** *64 × 1,024 = 65,536 bytes, which need 65,536 addresses*
   - **b)** *Each address represents one byte, so if there are only 256 possible addresses, there are only 256 bytes of memory. Hence, total size of memory = 256 × 8 = 2,048 bits*

4. *Answer for each part as follows:*
   - **a)** *False. The data bus connects the processor to memory only and not to secondary storage such as hard disks.*
   - **b)** *Two improvements (accept any possible answers):*
     - *Less vulnerable to damage from physical impact or external magnetic fields*
     - *Faster access speeds*
     - *Lighter weight*
     - *Less noise and power use due to no moving parts (longer laptop battery life)*